PATENT ABSTRACTS OF JAPAN

(11)Publication number : 02-254566
(43)Date of publication of application : 15.10.1990

(51)Int.Cl. G06F 15/401

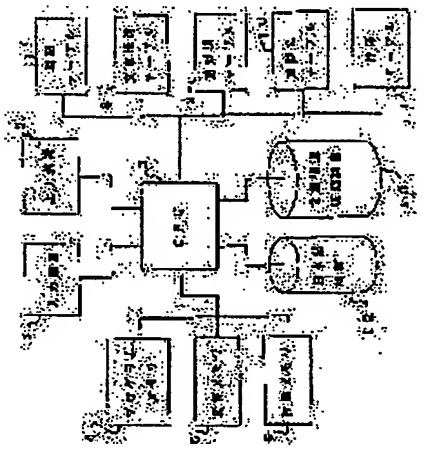
(21)Application number : 01-075214
(22)Date of filing : 29.03.1989

(71)Applicant : NIPPON TELEGR & TELEPH CORP <NTT>
(72)Inventor : NAGATA MASAAKI
KAWAI ATSUO
KIMOTO HARUO

(54) AUTOMATIC EXCERPT GENERATING DEVICE

(57)Abstract:

PURPOSE: To extract important words with a high precision to extract important sentences summarizing the points of contents by using information of a meaning paragraphic structure for extraction of important words and sentences to extract important words and sentences summarizing the points of respective meaning paragraphs and eliminating unnatural connection relations or the like from sentences including important word groups common to respective sentences to generates an excerpt.



CONSTITUTION: A processor 3 which executes an excerpt generating program, an input device 1 which reads in document data, and an output device 2 which outputs the generated excerpt to a magnetic storage device are provided. Further, an excerpt generating program memory 4, a document memory 5, a work memory 6, a word table 7, a sentence structure table 8, an important word table 9, an important sentence table 10, an excerpt table 11, a Japanese- language dictionary 12, and a sentence structure rule dictionary 13 are provided. The sentence structure of an original text is analyzed up to the level of meaning paragraphs to extract important words and sentences, and important word groups common to respective sentences are adopted and unnatural conjunctions and demonstratives are eliminated, and the sentence structure of the original text is used again to generate the excerpt which has unity and consistency with respect to sentence. Thus, important words are extracted with a high precision to extract important sentences summarizing the points.

LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

⑲日本国特許庁(JP)　⑪特許出願公開

⑫公開特許公報(A)　平2−254566

⑤1Int.Cl.[5]　識別記号　庁内整理番号　⑬公開　平成2年(1990)10月15日
G 06 F 15/401　7313−5B

審査請求　未請求　請求項の数 2　(全12頁)

⑤4発明の名称　自動抄録生成装置

⑳特　願　平1−75214
㉒出　願　平1(1989)3月29日

⑦2発明者　永田　昌明　東京都千代田区内幸町1丁目1番6号　日本電信電話株式会社内
⑦2発明者　河合　敦夫　東京都千代田区内幸町1丁目1番6号　日本電信電話株式会社内
⑦2発明者　木本　晴夫　東京都千代田区内幸町1丁目1番6号　日本電信電話株式会社内
⑦1出願人　日本電信電話株式会社　東京都千代田区内幸町1丁目1番6号
⑦4代理人　弁理士　金平　隆

明　細　書

1．発明の名称

自動抄録生成装置

2．特許請求の範囲

(1)　原文を入力するための入力部と、

形態素解析のための統語情報と意味情報を記憶した日本語辞書と、

この日本語辞書を用いて形態素解析を行う形態素解析部と、

見出し、段落、文等の文章の構成要素の間の関係に関する規則を記憶した文章構造規則辞書と、

この文章構造規則辞書、及び、名詞の使用状況に関する解析結果を用いて文書の構成要素を認識し、構成要素間の論理的な関係を解析する文章構造解析部と、

名詞の使用状況に関する解析結果を用いて、文章中の重要語を抽出する重要語抽出部と、

前記重要語抽出部より得られる重要語情報、及び前記文章構造解析部より得られる重要文について、その提示順序を決定し、抄録を生成する抄録生成部とを備え、

前記文章構造解析部は更に、

前記文章構造規則辞書を用いて、書式などの形態的特徴から文章の構成要素を認識する文章形態解析手段と、

この手段により認識された構成要素である段落について、名詞の使用状況を段落毎に解析し、連接する段落間の内容的関連性を判定して文章を区分化する段落連鎖解析手段とを備えたことを特徴とする自動抄録生成装置。

(2)　前記抄録生成部は、前記重要文抽出部により抽出された重要文を原文中の出現頻度に従って並べ、この重要文の系列において、原文の分脈とは整合しない接続詞及び指示詞を検出し、これを削除する手段を有することを特徴とする特許請求の範囲第1項記載の自動抄録生成装置。

3．発明の詳細な説明

[産業上の利用分野]

本発明は、文書データベース作成のために、

データベースに蓄積される文書に対して、重要な内容を簡潔に記述した抄録を、原文から自動的に生成する装置に関するものである。

［従来の技術］

マニュアル、新聞記事、特許出願の明細書、技術文献など、大量の文書から構成されるデータベースを作成する場合、文書の内容の概略を迅速に把握できるように、原文の抄録を作成することが必要である。従来、この目的のために、次のような方法が用いられていた。

（１）　何等かの文章の理解を行って、重要な文を決定する方法

（２）　文と文との関係を解析し、重要な文を決定する方法

（３）　キーワードの頻度を用いて、重要な文を決定する方法

しかし、これらの従来の方法には、それぞれ次のような欠点がある。

（１）の方法では、物語文法や因果関係を利用することにより、深層レベルの文章の構造的な制約や事象間の関係を解析し、これにより得られたデータ構造にに対して、要約規則を適用して重要な文を決定する。

この方法では、世界知識や推論規則など、対象に関する大量の知識と深い解析を行って抄録を生成する。このため、世界知識、常識、言語的な制約、意図、内容の関連、因果関係など、いろいろなレベルの知識を用いて、推論による論理的な帰結として、抄録を作成できるという利点がある。しかし、この方法は、非常に多くの世界知識と深い推論が必要なので、限られた狭い分野で、かつ、比較的短い文章にしか適用できないという欠点がある。

（２）の方法では、２つの文の連接関係（連続する２つの文の間の論理的な関係）の解析を行い、文の連接関係ごとに与えられた２つの文の相対的な重要度の指標に基づいて、原文中の文を取捨選択することにより抄録を生成する。この際、２つの文の間の連接関係は、接続詞と指示語、命題間の構成要素の概念関係、動詞・名詞・形容詞が持つ推論規則、スクリプトなどから求める。

この方法には、文と文の相対的重要度に関する規則から、推論による論理的帰結として抄録を導出できるという利点がある。しかし、多くの場合、文と文の関係は統語的な情報だけでは決定できず、多くの知識と推論を要するので、対象分野が非常に狭く限定されるという欠点がある。また、文の連接関係から得られる重要度は、段落内で最重要文を決定する場合のような局所的な重要度評価には有効であるが、文章全体の中から最重要文を決定する場合のような大局的な重要度評価には用いることができないので、比較的短い文章、または、文章の一部にしか適用できないという欠点がある。

（３）の方法では、文章の頻度統計などにより記述内容の主題や核となる重要語（キーワード）を予め求め、この重要語を多く含む文を重要文（キーセンテンス）として抽出することにより抄録を生成する。

この方法は、文章の大局的な解析により重要な内容を決定することができ、また、各文には重要語の頻度に応じて、重要度を付与することが出来るので、文章中から重要度の順に必要な数だけ重要文を選ぶことが出来るという利点がある。しかし、この方法には、抄録の中に文章の主題の展開とは余り関係のない文が混在したり、出力される抄録が互いに関連のない文の羅列となり文章としてのまとまりがない、等の欠点があった。

これらをまとめれば、（１）及び（２）の方法は、大量の知識を用いて非常に深い解析を行うので、抄録を作る対象が、限定された分野の短い文章に制限されるという問題点があった。一方、（３）の方法は、広い範囲の文章に対して適用可能であるが、文章主題の展開に関係のない文が抽出されたり、抄録に文章としてのまとまりがない、等の問題点があった。

［発明が解決しようとする課題］

本発明は、文章の展開に関係のない文が抽出されたり、抄録に文章としてのまとまりがない、という重要語の頻度に基づく抄録作成法の問題点を解決した自動抄録生成装置を提供することを目的

とする。

〔課題を解決するための手段〕

上記の課題を解決するために、本発明は、抄録生成プログラムを実行するプロセッサ、文書データを読み込む入力装置、生成された抄録を磁気記憶装置に出力するための出力装置、抄録生成プログラムメモリ、文書メモリ、作業メモリ、単語テーブル、文章構造テーブル、重要語テーブル、重要文テーブル、抄録テーブル、日本語辞書、文章構造規則辞書を備えている。

〔作用〕

本発明は、抄録作成処理の対象となる文章中で使用頻度が高い名詞は、主題や記述の核となる重要語であるという性質を利用して、プロセッサにおいて、日本語辞書を用いて機能語を完全に除去し、一般名詞と固有名詞を対象として、これらの頻度情報及び位置情報から、文章の主題や記述の核となる重要語を高精度で抽出する。

また、文章は、見出し、段落、文などを構成要素とする構造体であり、これらの構成要素間にはその構造を規定する規則が存在するという性質を利用して、構文解析の場合と同様に、統語的な手法により文章構造を解析する。

字下げで表現される形式段落の間には、内容的な結び付きの強いものと弱いものがあり、内容的関連度の高い一連の段落群を一つの意味段落としてまとめることができるという性質を利用して、各形式段落における名詞の使用状況を統計的に解析し、文章を意味段落に分割する。

また、重要語が初めて現れる文は、主題の導入や問題の提起を行う文であることが多いという性質、及び、重要語が最後に現れる文は、主題に関する結論を述べた文であることが多いという性質、段落の先頭にある文は段落の内容を総括する文であることが多いという性質、重要語を多く含む文は文章の中心的内容を述べていることが多いという性質を利用して、各文が含む重要語の頻度や文章内での文の位置から文章中の重要文を抽出する。

幾つかの意味段落から文章が構成されている場合には、中心的な話題を表現する重要語や、内容を総括する重要文が、意味段落ごとに存在することが多いという性質を利用して、意味段落ごとに重要語や重要文を抽出することにより、重要文抽出の際には、原文の文章構造における比較的大きな構成要素ごとに要旨を把握し、かつ、抄録生成の際には、原文における文章の展開と話題の変化を抄録に再現する。

同じ重要語群を含む文を集めると、文と文の間のつながり、すなわち、結束性が生じ、文章としての内容的なまとまりを持つ抄録となるという性質、及び、抽出された重要文を原文中の順番に従って並べ、原文と同様な文章構造を与えれば、原文内での文章の論理的構造及び情報の提示順序を抄録に再現することができることから、文章全体のまとまり、すなわち、一貫性を持つ抄録となるという性質を利用して、文章としての結束性と一貫性を兼備した抄録を生成する。

抽出された重要文を原文中の出現順序に並べると、これらの重要文中に含まれる接続詞及び指示詞の意味が、原文における意味と整合しないという性質と、連接する文間に共通の語群が存在すれば、接続詞や指示詞を除去しても結束性を維持できるという性質を利用して、抽出された重要文において、原文の文脈とは整合しない接続詞及び指示詞を検出し、これを削除することにより、文間の接続が自然で違和感がない抄録を生成する。

以上のような作用によって、本発明は、問題の提起、結論などの文章の展開を考慮し、主題と関係のない文が抄録に含まれることを防止し、原文の文章構造を意味段落のレベルまで解析して重要語や重要文の抽出に利用し、抄録生成の際に、各文に共通の重要語群を含ませ、不自然な接続詞や指示詞を除去し、原文の文章構造を再利用して文章としての結束性と一貫性を持った抄録を作成することを可能ならしめる。

〔実施例〕

第1図は、本発明の自動抄録生成装置を実現するための一実施例を示すシステムを構成する図である。1は磁気記憶装置に文字コードで記録されている文書データを読み込む入力装置、2は生成

された抄録を磁気記憶装置に出力するための出力装置、3は抄録生成プログラムを実行するプロセッサ（ＣＰＵ）、4は抄録生成プログラムを格納するプログラムメモリ、5は入力装置1により読み込まれた文書データを格納する文書メモリ、6は抄録生成プログラムを実行する際に使用する作業メモリ、7は形態素解析により得られた単語列を格納する単語テーブル、8は文章構造解析により得られた文章構造を記憶する文章構造テーブル、9は文章中から抽出された重要語を格納する重要語テーブル、10は文章中から抽出された重要文を格納する重要文テーブル、11は生成された抄録を格納する抄録テーブル、12は形態素解析及び名詞抽出を行う際に必要な統語情報と意味情報を格納した日本語辞書、13は処理対象とする文書の構造に関する規則を格納した文章構造規則辞書、である。各種テーブルは半導体メモリ、磁気ディスク、光ディスク等によって実現することができる。

第2図は、本発明の1実施例の機能ブロック図を示し、入力部21は第1図の入力装置1に相当するものであって、処理対象となる文書ファイル28を文書メモリ5に読み込む。

次に、形態素解析部22は、日本語辞書29を用いて、入力文章に対して辞書引きを行って単語単位に分割すると共に、各単語について品詞名を付与した単語列を生成し、単語テーブル7に格納する。この際、複合語（長単位名詞）は単位語に分割して格納する。

次に、文章構造解析部23と重要語抽出部24の処理が並行して実行され、解析結果を相互に参照しながら、文章構造の情報と重要語の情報を重要文抽出部に引き渡す。

文章構造解析部23は、第3図に示すように、書式から認識可能な文書構造を生成する文章形態解析手段（ステップｓ1）と、形式段落間の内容的関連度を解析して文章を意味段落に分割する段落連鎖解析手段（ステップｓ2）から構成され、文章の構成要素（文、段落、見出し、意味段落など）及び構成要素間の関係を文章構造テーブル8に格納する。以下に、文章解析部23の各ステップを詳細に説明する。

ステップｓ1の文章形態解析では、入力文章を文字列と見なし、この文字列に対して、文章構造規則辞書30に格納されている文章構造規則を適用して、まず、題名、見出し、文などの文章の基本的な構成要素を認識する。次に、これらの構成要素の間の関係を解析して、段落や節などのより大きな文章の構成要素を認識する。

文章構造規則は、「文書」を初期記号とし、印字可能文字と書式制御文字を終端記号とする生成規則群とする。非終端記号としては、「見出し」、「段落」、「文」などが用いられる。多くの場合、文章構造の形式文法による記述は、文脈自由文法（ＣＦＧ）の枠内で行うことができる。この規則を利用して、文章形態解析では、トップダウンＣＦＧパーサにより文章構造を解析する。

第4図は、文章構造規則を用いた文章構造に関する形式文法の定義の一例である。規則の第1行は、この例で処理対象としている文書は、文書見出し部は、題目、所属、著者から構成されることを表し、第6行から第9行は、文書本体部は節の繰り返しであり、節は節見出し部を持つ場合と持たない場合があり、節本体部は、段落の繰り返しであることを表している。これらの規則は、対象とする文書に固有なものであり、規則を書き換えることにより、様々な文書に対して、文章形態解析を適用することができる。

文章形態解析（ステップｓ1）は、最後に、解析の結果として得られた、題名、段落、文等に関する文章構造情報を文章構造テーブル8に格納し、処理を終了する　ステップｓ2の段落連鎖解析では、各段落（形式段落）ごとに計算した語彙的な特徴量を時系列分析することにより段落間の内容的な関連度を調べ、文章を幾つかの意味的なまとまり（意味段落）に区分化する。

第5図は、段落連鎖解析手段のフローチャートである。段落連鎖解析では、まず、各段落ごとに名詞を抽出し、次のような項目から構成される段落語彙表を作成する（ステップｓ11）。

①単位語集合（同語反復を含む全ての名詞の集合）

②見出し語集合（全ての異なり名詞の集合）

③延べ語数（単位語集合の要素数）

④異なり語数（見出し語集合の要素数）

次に、文章を意味段落に分けるために、内容的な結び付きの強さを指標化するための特徴量を各段落（形式段落）ごとに計算する（ステップｓ１２）。

ここでは、文章の展開に伴う各段落中の異なり語数と延べ語数の変化、及び、隣接する段落間の同一語句の反復という現象に着目し、段落間の内容的な結び付きの強さの指標として、次のような新出語率と用語類似度という２つの統計的な尺度を定義する。

$$\text{新出語率 } r(i) = \frac{\Delta k(i)}{\Delta n(i)} = \frac{k(i) - k(i+1)}{n(i) - n(i+1)}$$

k（ｉ）：段落ｉに初めて現れる語の異なり語数

n（ｉ）：段落ｉの延べ語数

$$\text{用語類似度 } D(i) = (1/N(i)) \sum_{M \in V(i+1)} F(i, M)$$

V（ｉ）：段落ｉ上の語彙（異なり語の集合）

N（ｉ）：段落ｉの延べ語数

M：見出し語（異なり語）

F（ｉ，Ｍ）：段落ｉ中のＭの使用度数

新出語率ｒ（ｉ）は段落ｉでの見出し語の平均増加率である。話題の変化点では新語が群出する傾向があるので、新出語率の極大点から、新しい話題の導入点を検出できる。

また、用語類似度Ｄ（ｉ）は、段落ｉとその直前の段落ｉ－１で、共通の語が用いられていることが多いので、用語類似度の極小点から話題の変化点を検出できる。

文章を段落の時系列として捉えると、一般に、文章の進展に伴って、新出語率は減少し、用語類似度は漸増する傾向がある。これは、文章の終わりに近づくにしたがって、話題が結論へと徐々に収束して行くためである。このような傾向変動を考慮し、大局的観点から話題の変化点を検出するために、まず、次のようにして、段落の時系列における新出語率と用語類似度の傾向変動を最小２乗法により１次関数で近似する（ステップｓ１３）。

一般に、系列Ｘ（ｉ）が傾向変動Ｔ（ｉ）を含み、それ以外を残差系列ｅ（ｉ）として

$$X(i) = T(i) + e(i)$$

と表すとき、傾向変動Ｔ（ｉ）を１次関数

$$T(i) = \beta_0 + \beta_1 * i$$

で近似することにすれば、最小２乗法により、ｅ（ｉ）の２乗和Ｑ

$$Q = \Sigma \{e(i)\}^2 = \Sigma \{X(i) - (\beta_0 + \beta_1 * i)\}^2$$

を最小にするようなβ_0及びβ_1を決めることができる。これより、残差系列ｅ（ｉ）は、

$$e(i) = X(i) - T(i) = X(i) - (\beta_0 + \beta_1 * i)$$

として求めることができる。

この方法を新出語率　ｒ（ｉ）及び用語類似度Ｄ（ｉ）に適用することにより、新出語率の残差$e_r(i)$及び用語類似度の残差$e_D(i)$を求める。

次に、残差系列が極値（新出語率は極大値、用語類似度は極小値）をとる段落を意味段落の切れ目の候補として選択する（ステップｓ１４）。具体的には、次の条件のどちらか一方を満足する段落ｉを意味段落分割点の候補とする。

(a) $e_r(i) > 0, \quad e_r(i-1) < e_r(i), \quad e_r(i) > e_r(i+1)$

(b) $e_D(i) < 0, \quad e_D(i-1) > e_D(i), \quad e_D(i) < e_D(i+1)$

次に、この候補に基づいて、最終的に文章全体がほぼ同じ大きさの２～４つの部分に分かれるように意味段落を決定する（ステップｓ１５）。

具体的な手順は次の通りである。

①　分割点候補を

$\max(|e_r(i)|, |e_D(i)|)$

の降順に整列する。

②　隣接する段落が両方とも分割点候補となっているときは、

$\max(|e_r(i)|, |e_D(i)|)$

の小さい方を候補から除去する。

③　$\max(|e_r(i)|, |e_D(i)|)$

の大きい方から決められた数だけ分割点を

選ぶ。

但し、元の文章が、文章形態解析の段階で幾つかの部分に分けられているときは、文章形態解析の結果、すなわち、書式などの解析により得られた章節構造を優先する。

文章形態解析は、最後に、意味段落の情報を文章構造テーブルに格納し、処理を終了する。

第6図は、重要語抽出部24のフローチャートである。重要語抽出部では、まず、ステップs21において、単語テーブル7上の単語の品詞情報を参照して、名詞以外の単語、及び、時詞、数詞、代名詞、形式名詞など機能語的な役割を持った名詞を除去し、一般名詞及び固有名詞のみを抽出する。

つぎに、ステップs22において、語彙の頻度統計を行い、続いて、ステップs23において、見出し語集合を次のような情報をキーとして整列する。

第1キー：使用度数（降順）

第2キー：見出しに出現する／しない（出現する方を先に並べる）

第3キー：初出位置（昇順）

整列された各見出し語について、次のような統計量からなる度数順単語表を作成する。

① 使用度数（頻度）の順位

② 単語の字面（見出し語）

③ 使用度数

④ 累積使用率（延べ語被覆率）

⑤ 異なり語被覆率

⑥ 見出し出現語フラグ

⑦ 初出位置

一般に、使用頻度の高い単語は、重要語となる確率が高い。しかし、高頻度語の集合による重要語の被覆率を調べると、重要語の被覆率は、高頻度語集合の大きさ（異なり語の被覆率）の対数と比例する。すなわち、頻度の非常に高いグループ語が重要語に選ばれる確率は非常に高いが、頻度が低くなるに従って、異なり語数が指数的に増えるので、次に頻度が高いグループの語が重要語に選ばれる確率は、前のグループよりかなり低くなる。

そこで、重要語集合の中で、重要度の順位付けをする必要が生じたときのために、第1のグループの語を最重要語、第1と第2のグループの語を合わせたものを重要語と呼ぶことにし、以降の処理では扱いを別にする。

以上の知見に基づいて、ステップs24では、累積使用率が15%以下の語を最重要語として抽出する。また、ステップs25では、異なり語被覆率が5%を超えない範囲で、累積使用率が25%以下の語を重要語として抽出する。なお、このしきい値は、専門家と同程度の量の重要語が得られるように、標本として用いた文章から実験的に決定したものである。対象とする文章の性質及び生成しようとする抄録の性質に応じて、しきい値を変更することにより、本手法は各種の文章に対して適用可能である。

重要語抽出部24は、最後に、抽出した最重要語及び重要語は、重要語テーブル9に格納し、処理を終了する。

第7図は、重要文抽出部のフローチャートである。重要文抽出部25は、文章構造テーブル8及び重要語テーブル9の情報に基づいて、まず、原文中の各文の形態的な特徴を解析する（ステップs31）。

文特徴解析（ステップs31）では、重要語抽出の結果と文章構造解析の結果に基づいて、次の3つの観点から文の特徴を指標化する。

（a）各文の最重要語及び重要語の頻度

（b）文の時系列上での最重要語及び重要語の分布

（c）文の構造体の中での文の位置

また、大局的な文脈を特徴量に反映させるために、文章構造解析の結果に基づいて、次の2種類の重要語を用意する。

（i）文章全体の最重要語及び重要語

（ii）各意味段落の最重要語及び重要語

これらの観点から、文特徴解析（ステップs31）では、各文について次のような項目からなる文特徴表を作成する。

① 意味段落番号、段落番号、段落内文番号
② 文章全体の中で、その文で初めて出現した最重要語
③ 文章全体の中で、その文で最後に出現した最重要語
④ その文に出現する文章全体の最重要語及び重要語
⑤ その文に出現する意味段落の重要語

次に、ステップｓ３２では、文特徴解析の結果から文の重要度の順位付けを行う。重要語を含む文は、文章の主題のある側面を記述していることは明らかである。しかし、意味的な情報を使わない場合には、何が重要な側面であるかは、確率的な尺度で決めざるを得ない。

内容抽出の観点から考えられる重要文の条件は次のようなものである。

（ａ）重要語を多く含む文
（ｂ）重要語の初出文及び終出文
（ｃ）段落の先頭にある文

重要語を多く含む文は、文章の中心的内容を述べていることが多く、その文から原文中の多くの内容が連想可能である。重要語の初出文や終出文は、問題の提起、結論などを多く含むことが多い。また、段落の先頭にある文は段落の内容を総括する文であることが多い。

文章生成の観点から考えられる重要文の条件は次のようなものである。

（ｄ）各文が文章の主題に関連する語群を含む
（ｅ）主題の導入的記述を含む文

複数の文により内容を記述する場合、各文の間の結束性と文章自身の完結性が要求される。結束性を与えるためには、各文が、同じ主題に対する叙述でなければならない。また、完結性を与えるためには、読み手が主題を認識できる記述を抄録中に含むことが必要である。

これらの考察から、次の①②③④の順に（④は、重要語が多い順に）、文の重要度の順位付けを行う。

① 最重要語が初めて出現する文
② 最重要語が最後に出現する文
③ 最重要語を含み、段落の先頭にある文
④ 最重要語を含み、かつ、重要語が多く出現する文

次に、ステップｓ３３では、重要度の高い文から順に、文数が原文の２０％に達するまで文を選ぶ。ここで、しきい値は、標本として用いた文章において、専門家が選ぶ重要文の数と同程度の量の重要文が、自動抄録により得られるように実験的に決定した。対象文章及び生成すべき抄録の性質に応じてしきい値を変更することにより、本手法を各種の文章に適用することができる。

更に、意味段落の情報が利用可能な場合には、①②の重要文は文章全体から抽出し、③④の重要文は意味段落ごとに抽出する。この際、最重要語は、文章全体の重要語と各意味段落の重要語を併合したものを用いる。

重要語抽出部２５は、最後に、こうして決定された重要文を重要文テーブル１０に格納し、処理を終了する。

第８図は、抄録生成部のフローチャートである。抄録生成部２６は、まず、ステップｓ４１において、選ばれた重要文を原文中の順に並べ直す。これは、原文中の文の順序を反映し、抄録を原文の類似縮小形とすることにより、文章としての一貫性を持つ抄録を生成するためである。

次に、ステップｓ４２では、この文の列の中で、文頭に位置する接続詞や指示詞を単語テーブル７の品詞情報を用いて検出する。これは、接続詞や指示詞は、原文の文脈において文間の関係を示していた語なので、重要文の列の中では原文の文脈と整合しない接続関係や参照関係を生じさせる可能性を持っているためである。

次に、ステップｓ４３では、ステップｓ４２で検出した接続詞及び指示詞の中で、原文の文脈と整合しない抄録中の接続詞及び指示詞を削除する。原文の文脈と整合するかどうかの判断は、重要文の原文での位置情報を用いて、次のように行う。

一般に、原文中で連続する２つの文が重要文に選ばれた場合には、前方の文に対する後方の文の接続関係及び参照関係が抄録中でも保持される可

能性が非常に高い。また、文頭の接続詞や指示詞を除去しても文が非文（文法的に正しくない文）になることはない。さらに、複数の文が共通の語群を含んでいれば、文間に結束性が生じるので、文頭の接続詞や指示詞がなくても文間の接続関係や参照関係は、ある程度まで読者が保管することができる。

これらを考慮して、ステップｓ４３では、原文中で連続する２つの重要文における後方の文の接続詞及び指示詞は保存し、それ以外は全て削除する。

次に、ステップｓ４４では、生成すべき抄録の書式に従って、題名、著者名、及び、抄録を書式付けし、その結果を抄録テーブル１１に格納し、抄録生成部２６の処理を終了する。

最後に、出力部２７は、抄録テーブル１１の内容を抄録ファイル３１に格納する。

第９図は、本発明の一動作例である。

入力文書の原文を第９図ｅに示す。入力部２１により読み込まれた文書ファイル２８中の入力文書は、形態素解析部２２により形態素解析される。

次に、文章構造解析部２３において、まず、文章形態解析が文章構造規則辞書３０中の文章構造規則を用いて、入力文書の構造解析を行う。この結果を第９図ａに示す。この処理により、入力文書における、題名、段落、文などの各構成要素を認識し、さらに文と段落の関係など構成要素間の関係も認識する。

続いて、段落連鎖解析が各段落の用語類似度及び新出語率を計算した結果に基づいて、意味段落分割を行う。用語類似度と新出語率の計算結果、及び、得られた意味段落を第９図ｂに示す。この例では、新出語率の誤差系列の極大点から段落４と段落１３が、また、用語類似度の誤差系列の極小点から段落９が大局的にみた話題の変化点として選ばれ、意味段落が決定される。

次に、重要語抽出部２４は、語彙統計により度数順単語表を作成し、文章全体の最重要語と重要語、及び、各意味段落の重要語を決定する。文章全体の度数順単語表、文章全体の最重要語と重要語、及び、各意味段落の重要語の抽出結果を第９図ｃに示す。この例では、累積使用率１５％以下の「競争」、「サービス」、「技術」、「仕事」、「ＡＳＫ」が最重要語として選ばれ、これらに加えて、累積使用率２５％以下という条件を満足する「自分」、「活動」、「失敗」などが重要語として選ばれる。また、各意味段落での語彙統計により、意味段落ごとの重要語も同様に決定される。

次に、重要部抽出部２５は、各文の文特徴解析を行う。文特徴解析の結果を第９図ｄに示す。文特徴解析では、各文に対して、その文の位置、字面、初出及び終出する最重要語及び重要語、文中の最重要語、文中の意味段落重要語などに関する情報を付与する。続いて、文解析の結果に基づいて、文を重要度の順に整列し、必要な分量の文を選択する。重要文抽出の結果を第９図ｅに示す。この例では、原文の２０％を超えるまで文を選ぶという条件により、原文中の４８個の文の中から１０個の文が抽出される。

次に、抄録生成部２６は、抽出された重要文を原文の順番に並べ直し、文頭の接続詞し及び指示詞の中で、不適切なものを除去する。削除すべき接続詞及び指示詞を検出した結果を第９図ｆに示す。この例では、「しかし」、「ですから」、「これが」、「しかし」の４つが文頭の接続詞・指示詞として検出されるが、後者の２つの語は原文中の直前の文が抄録にも選ばれているので削除されず、前者の２つの語が削除される。

最後に、出力部２７は、出来上がった抄録を抄録ファイル３１に書き出す。

〔発明の効果〕

上記の説明のように、本発明は、

日本語辞書の品詞情報と、単語の頻度情報、及び位置情報から文章の主題や記述の核となる重要語を高精度で抽出し、

問題の提起や結論などの原文の文章の展開に基づいて、文章展開の上で重要であり、かつ、内容の要点を述べた重要文を抽出し、

原文の文章構造を意味段落のレベルまで解析することにより、意味段落構造の情報を重要語や重

要文の抽出に利用して、意味段落ごとの要旨を述べた重要語や重要文を抽出し、

また、抄録作成の際には、各文に共通の重要語群を含んだ文を、原文の論理的構造を利用し、不自然な接続関係や参照関係を除去して抄録を生成するものであるから、

文章の主題や記述の核となる重要語を構成度に抽出することができ、

文章展開上で重要であり、かつ、内容の要点を述べた重要文を抽出することができ、

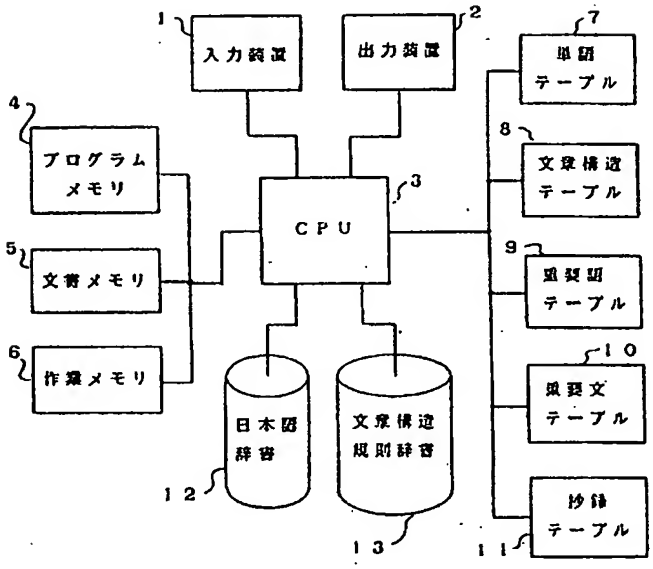
また、文章としての結束性と一貫性を持った抄録を生成することができる、という改善効果が得られた。

4．図面の簡単な説明

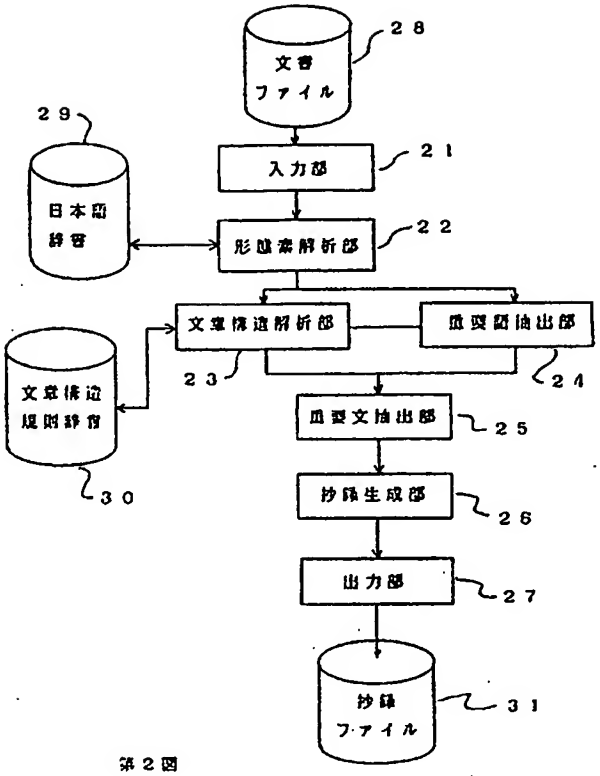
第1図は本発明の1実施例のシステム構成図、第2図は本発明の1実施例の機能ブロック図、第3図は文章構造解析部のフローチャート、第4図は文章形態解析で用いる形式文法及び文章構造規則の1例、第5図は段落連結解析のフローチャート、第6図は重要語抽出部のフローチャート、第7図は重要文抽出部のフローチャート、第8図は抄録生成部のフローチャート、第9図は本発明の1動作例である。

1：入力装置
2：出力装置
3：プロセッサ（ＣＰＵ）
4：プログラムメモリ
5：文書メモリ
6：作業メモリ
7：単語テーブル
8：文章構造テーブル
9：重要語テーブル
10：重要文テーブル
11：抄録テーブル
12：日本語辞書
13：文章構造規則辞書
21：入力部
22：形態素解析部
23：文章構造解析部
24：重要語抽出部
25：重要文抽出部
26：抄録生成部
27：出力部
28：文書ファイル
29：日本語辞書
30：文章構造規則辞書
31：抄録ファイル

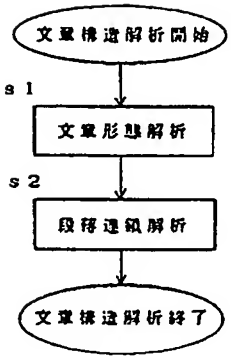
出願人　日本電信電話株式会社
代理人　弁理士　金　平　　隆



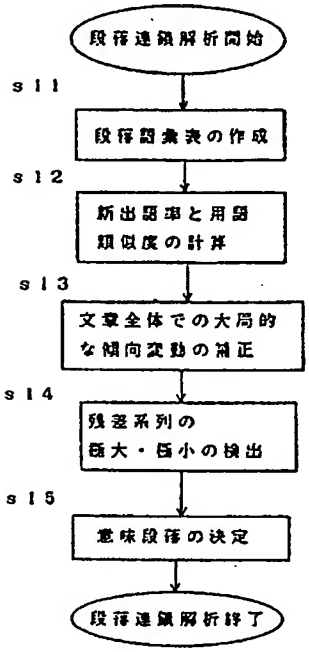
第1図



第2図



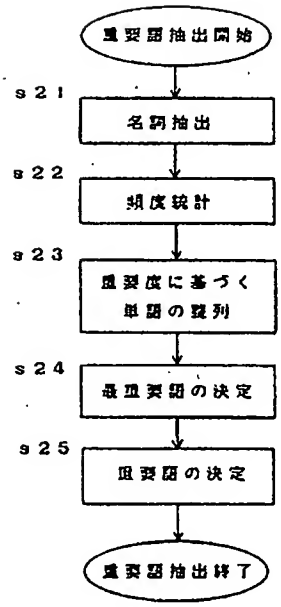
第3図



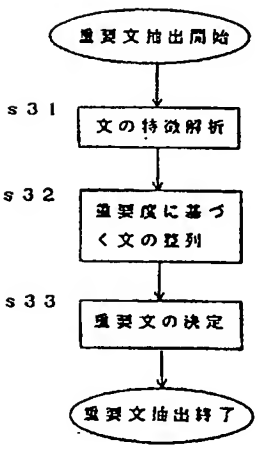
第5図

```
一般的な形式文法を　G＝（Vn，Vt，P，S）
　　Vn：　非終端記号の集合
　　Vt：　終端記号の集合
　　P：　生成規則の有限集合
　　S：　初期記号
と表すことにする。


　　文章構造に関する形式文法の例

G1＝（Vn，Vt，P，S）
  Vn＝｛＜文書見出し部＞，＜文書本体部＞，...｝
  Vt＝｛α｜　印字可能文字，または，書式制御文字｝
  S＝｛＜文書＞｝
  P＝｛
　　＜文書＞::＝＜文書見出し部＞＜文書本体部＞
　　＜文書見出し部＞::＝＜題目＞＜所属＞＜著者＞
　　＜題目＞::＝＜文＞
　　＜所属＞::＝＜文＞
　　＜著者＞::＝＜文＞
　　＜文書本体部＞::＝｛＜節＞｝
　　＜節＞::＝＜節見出し部＞＜節＞｜＜節本体部＞
　　＜節見出し部＞::＝＜文＞
　　＜節本体部＞::＝｛＜段落＞｝
　　＜段落＞::＝＜段落開始記号＞｛＜文＞｝
　　＜文＞::＝｛＜文字＞｝＜文末記号＞
　　＜段落開始記号＞::＝＜空白文字＞
　　＜文末記号＞::＝句点『。』｜疑問符『？』
　　　　｜感嘆符『！』｜＜空白文字＞
　　＜文字＞::＝＜文末記号以外の印字可能文字＞
　　＜空白文字＞::＝空白｜タブ｜改行｜改頁
  ｝
```

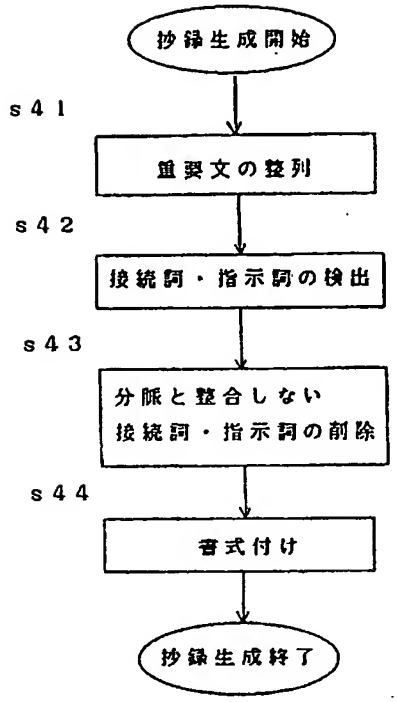
第4図



第6図



第7図

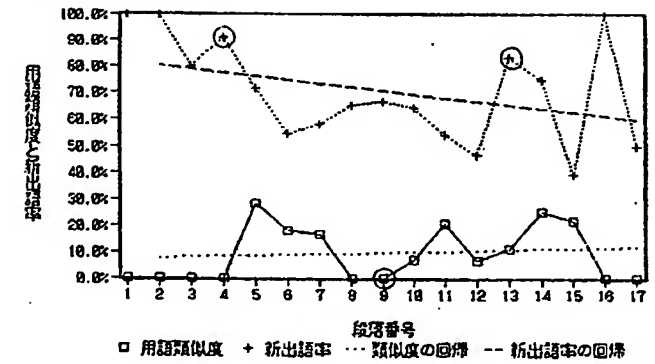


第8図

```
(文書
  (文書見出し部
    (題名 (文 "年頭に当たって"))
  (文書本体部
    (節
      (節本体部
        (段落
          (文 "皆さん新年明けましておめでとうございます。"))
        (段落
          (文 "昨年は東証一郎への. . . 注目された年でした。")
          (文 "そうした中で. . . 行うことができました。"))
        (段落
          (文 "これは、あなたたちが. . . 結果です。")
    *****  中略  *****
        (段落
          (文 "今年は" 本格的な競争時代" . . . 歴然としています。")
          (文 "私のもとに来た手紙も. . . もらいました。")
          (文 "皆さんの健康と活躍を祈っています。"))))))
```

第9図 (a) 文章形態解析

〔用語類似度と新出語率〕



〔意味段落〕
段落 1 ～ 段落 3
段落 4 ～ 段落 8
段落 9 ～ 段落12
段落13 ～ 段落17

第9図 (b) 段落連続解析

〔文章全体の度数順単語表〕 延べ語数 251 異なり語数 173

順位	単語	使用度数	累積使用率	異なり語被覆率	見出し-F	初出位置
1	競争	10	4.0%	0.6%	0	(6 1)
2	サービス	8	7.2%	1.2%	0	(8 1)
3	技術	7	10.0%	1.7%	0	(8 1)
4	仕事	5	13.9%	2.3%	0	(8 1)
4	ＡＳＫ	5	13.9%	2.9%	0	(14 4)
6	自分	4	16.7%	3.5%	0	(4 2)
6	活動	4	16.7%	4.0%	0	(14 4)
6	失敗	4	16.7%	4.6%	0	(17 1)
9	いま	3	27.9%	5.2%	0	(3 1)
9	企業	3	27.9%	5.7%	0	(5 1)

***** 以下省略 *****

〔文章全体の最重要語〕
"競争","サービス","技術","仕事","ＡＳＫ"

〔文章全体の重要語〕
"競争","サービス","技術","仕事","ＡＳＫ","自分""活動","失敗"

〔各意味段落の重要語〕

意味段落番号	段落番号	意味段落内の重要語
1	1～ 3	"企業"
2	4～8	"技術","サービス"
3	9～12	"競争","仕事"
4	13～17	"失敗"

第9図 (c) 重要語抽出

```
      *****   前半省略   *****

(文
      :意味段落番号 4 :段落番号 13 :段落内文番号 2
      :字面 "現場であろうが、技術系であろうが、営業系であろうが、ＡＳＫ
活動で発想を変え、合意に基づき実行して更に次の段階に進歩させていく。"
      :切出最重要語 NIL :終出最重要語 ("技術")
      :最重要語 (("技術" 1) ("ＡＳＫ" 1))
      :重要語 (("技術" 1) ("ＡＳＫ" 1) ("活動" 1))
      :意味段落重要語 NIL)

(文
      :意味段落番号 4 :段落番号 17 :段落内文番号 1
      :字面 "ＡＳＫ活動は、繰り返し繰り返しやっていけば、時として失敗も
あるでしょう。"
      :切出最重要語 NIL :終出最重要語 NIL
      :最重要語 (("ＡＳＫ" 1))
      :重要語 (("ＡＳＫ" 1) ("失敗" 1) ("活動" 1))
      :意味段落重要語 ("失敗" 1))

(文
      :意味段落番号 17 :段落番号 17 :段落内文番号 2
      :字面 "しかし、このＡＳＫの過程で出てきた失敗は、自分で考えて実行
した本人が一番先に見つけます。"
      :切出最重要語 NIL :終出最重要語 ("ＡＳＫ")
      :最重要語 (("ＡＳＫ" 1))
      :重要語 (("ＡＳＫ" 1) ("失敗" 1) ("自分" 1))
      :意味段落重要語 ("失敗" 1))

      *****   後半省略   *****
```

第９図 （ｄ）重要文抽出

年頭にあたって

①皆さん、新年明けましておめでとうございます。

②昨年は現証１部への上場や政府保有株式の２回目の放出など、我が社としては民営化移行時以上に世間に注目された年でした。そうした中で自分たちの本分を見失わず、経営の効率化や事業領域の拡大などを行ってきた結果、業績のほうも順調に推移し、公益事業体として恥ずかしくない１割配当を行うことができました。

③これは、あなたたちが一致協力し、企業人として努力してきた結果です。一歩一歩着実に"民間企業"としての体質が整ってきているな、と私自身実感しており、社長として、心から「ありがとう」と申し上げたいと思います。

④しかし、そうした社内の状況から世の中の動向に目を移しますと、従来からのデータ通信やＶＡＮ、専用線、端末機などの分野に加えて、昨年後半からは市外電話やポケットベルの分野でも"競争"が始まっています。あなたたちの本当の実力が試されるのは"今からなのだ"ということをしっかりと肝に銘じてほしいと思います。

⑤この競争の世の中でＮＴＴの皆の我々仲間にとって大切なのは、まず、技術開発の力です。

⑥ＮＴＴのように技術力をベースにして、精度の高い設備を駆使してサービスを提供する仕事の場合には、競争のポイントを、まず、技術開発力に置かなくてはいけません。技術力をベースにするということは、新しいサービスを提供するための新しいシステムを考案する「開発の力」と、それを最も効率的に据え付け、実際のサービスに最も合理的に機能させる「運用技術の開発」を進めていくということです。そのことがすべての基本になります。

⑦次に、具体的に安全で故障の少ない設備を構築し、"お客様に迷惑をかけない"ようにすることです。そのためには、サービスのあり方を日々新たに改良し、それが何とはなしにお客様に分かって頂けるところまでもっていってもらいたいと思います。

⑧我々は技術的な観点からどれだけ実現可能なものであっても、一般の人にとって必ずしも高い価値を持つとは限らない、ということを認識しなくてはいけません。Ｒ＆Ｄに当てはめて言えば、私達は現在の社会のニーズを正確に把握し、将来のニーズを正確に予測してＲ＆Ｄを進めなければいけません。その意味で、昨年７月にＲ＆Ｄ体制の見直しを行いましたが、今後とも人と組織の両面から一層充実した開発体制を整える必要があります。

⑨このような総合的な技術力を持つためには、その基礎となる基盤技術レベルの高さが必要なことはもちろんです。

⑩次に財政基盤の確立です。競争相手との競争の過度状況を把握して、やはりこういうサービスの場合にはもう少し値下げしなければいけないな、と思ったときにいちでも値下げできるような財務体質の強さを持つことです。

⑪競争というのは突き詰めると料金の"値下げ"競争なのです。それが、非常に高度な技術を必要とするサービスであれば、我々に「分」があるのは確かなのですが、現在の収入の大部分は普通の電話からの料金で回っているのが実情です。したがって、料金競争を普通の電話料金の分野で行うとして、仮に競争に勝ってシェアは取っても、我々自身が赤字になることも考えられます。

⑫ですから、私が皆さんにお願いしたいことは、今までの仕事のやり方を更に物理的に改善し、より効率化を図って、より安いサービスが提供できるように知恵を出してもらいたいということです。仕事のやり方を日々新しく改良していく努力を、あなたたち一人ひとりが考えてみんなと相談すれば、今までよりも仕事がしやすく、より安くできます。一人ひとりが仕事をしやすいということは、安くできるということと同じことなのです。これが皆さんにお願いしているＡＳＫ活動そのものです。

⑬単に、経費を節約すればいいんだ、前例に従ってオートマチックに動けばいいんだ、という考えではダメです。現場であろうが、技術系であろうが、営業系であろうが、ＡＳＫ活動で発想を変え、合意に基づき実行して更に次の段階に進歩させていく。この過程を繰り返していくことが、進歩そのものです。このことは６０年近く技術畑で暮してきた私の生活そのものだと自信をもって申し上げます。

⑭昨年、私はあなたたちに「創下精励」という言葉を送りました。足下を見つめてほしい、つまりＡＳＫ活動の原点をみつけてほしい、という願いを込めてのことです。

⑮ＡＳＫ活動は、繰り返し繰り返しやっていけば、時として失敗もあるでしょう。しかし、このＡＳＫの過程で出てきた失敗は、自分で考えて実行した本人が一番先に見つけます。この自分でまずいところを見出だすことができれば、その欠点の修正方法は半ば本能的に分かります。失敗を自分で見つける嗅力が自分の実力の進歩そのものです。創下精励の力です。

⑯余り考えると結核になります。純粋に、無心の態度で考えていれば、結核になるどころか逆に病気が出てきます。「千万人といえども吾往かん」という心境がこれです。

⑰今年は"本格的な競争時代"が始まったということもあって、お正月早々お祝気じみてしまいましたが、あなたたちが企業人として着実に進歩していることは歴然としています。私のもとに来た手紙も、随分と考え方が変わってきたなと思いながら読ませてもらいました。今年も失敗を恐れず、果敢に挑戦し、お客さまに喜ばれるサービスに努めてほしいと思います。皆さんの健康と活躍を祈っています。

①②③：段落番号
＿＿＿：自動抄録の重要文

第９図 （ｅ）重要文抽出の結果 2/2

年頭にあたって

しかし、そうした社内の状況から世の中の動向に目を移しますと、従来からのデータ通信やＶＡＮ、専用線、端末機などの分野に加えて、昨年後半からは市外電話やポケットベルの分野でも"競争"が始まっています。ＮＴＴのように技術力をベースにして、精度の高い設備を駆使してサービスを提供する仕事の場合には、競争のポイントを、まず、技術開発力に置かなくてはいけません。ですから、私が皆さんにお願いしたいことは、今までの仕事のやり方を更に物理的に改善し、より効率化を図って、より安いサービスが提供できるように知恵を出してもらいたいということです。一人ひとりが仕事をしやすいということは、安くできるということと同じことなのです。これが皆さんにお願いしているＡＳＫ活動そのものです。現場であろうが、技術系であろうが、営業系であろうが、ＡＳＫ活動で発想を変え、合意に基づき実行して更に次の段階に進歩させていく。ＡＳＫ活動は、繰り返し繰り返しやっていけば、時として失敗もあるでしょう。しかし、このＡＳＫの過程で出てきた失敗は、自分で考えて実行した本人が一番先に見つけます。今年は"本格的な競争時代"が始まったということもあって、お正月早々お祝気じみてしまいましたが、あなたたちが企業人として着実に進歩していることは歴然としています。今年も失敗を恐れず、果敢に挑戦し、お客さまに喜ばれるサービスに努めてほしいと思います。

（ＮＴＴ施設,Vol.40 No.1,1988）

＿＿＿：削除すべき接続詞及び指示詞
.....：削除しなくてもよい接続詞及び指示詞

【圧縮率】
原文 -> 文数： 48
抄録 -> 文数： 10
圧縮率: 20.8 %

第９図 （ｆ）抄録生成